



现代语言测试与评估丛书

曾用强 总主编

计算机技术、语料库 与语言测试

王永庆 著



科学出版社

现代语言测试与评估丛书

曾用强 总主编

计算机技术、语料库 与语言测试

王永庆 著



科学出版社

图书在版编目(CIP)数据

计算机技术、语料库与语言测试 / 王永庆著. —北京:
科学出版社, 2014. 1

(现代语言测试与评估丛书 / 曾用强主编)

ISBN 978-7-03-039384-5

I. ①计… II. ①王… III. ①数理语言学-语料库-应用-
语言-测试-研究 IV. ①H09

中国版本图书馆 CIP 数据核字(2013)第 309819 号

责任编辑: 刘彦慧 张翠霞 / 责任校对: 何艳萍

责任印制: 钱玉芬 / 封面设计: 无极书装

联系电话: 010-64019074 电子邮箱: liuyanhui@mail.sciencep.com

科学出版社出版

北京东黄城根北街 16 号

邮政编码: 100717

<http://www.sciencep.com>

双青印刷厂印刷

科学出版社发行 各地新华书店经销

*

2014 年 2 月第 一 版 开本: A5 (890×1240)

2014 年 2 月第一次印刷 印张: 6 1/4

字数: 200 000

定价: 58.00 元

(如有印装质量问题, 我社负责调换)

总 序

据 Spolsky 考证,正式的语言测试起源于中国东汉时期的科举考试。但是,现代意义上的语言测试却于 20 世纪中叶诞生在英美等发达国家。20 世纪 60 年代,外语测试作为一门新的学科从外语教学中独立出来。作为语言学、教育与心理测量学、计算机技术等交叉学科,半个世纪以来,国外大批语言学家在语言测试领域取得了卓越的成就。两本专业期刊——《语言测试》(*Language Testing*)和《语言评估季刊》(*Language Assessment Quarterly*)相继诞生,在应用语言学界的影响越来越大,语言测试已经成为应用语言学领域的显学之一。在我国这样一个具有浓郁考试文化的大国,考试一直备受国人青睐,甚至受到顶礼膜拜。中国的高考和大学英语考试(CET)虽有数百万考生规模,但在研究方面仍落后于英美等发达国家。至今,以美国 ETS(Educational Testing Service)开发的 TOEFL 和英国剑桥大学考试委员会(CESOL)的 IELTS 为代表的西方国家的研究水平仍执全球之牛耳。但是,我们欣喜地看到,国内学者没有妄自菲薄,外语测试研究日益受到人们的重视,并已经取得一些成绩,一些学者的论文在国际期刊发表,国内专业期刊《外语测试与教学》也在 2011 年于上海外国语大学问世。虽然有若干著述出版,但它们大多关注于测试的开发实践,而对测试理论与实践的研究较少。在此背景下,我和我的博士们策划了“现代语言测试与评估丛书”。在国内学界享有盛誉的科学出版社高瞻远瞩,大力扶持外语研究,欣然同意出版这套丛书,

可谓语言测试界的盛举。丛书为国内语言测试研究者提供一个平台，系统展示国内外语言测试领域的新成果，特别是国内学者的原创研究，供广大同行分享。

本丛书分为两个系列：①语言测试与评估：研究系列；②语言测试与评估：实践系列。前者以理论和研究为重点，主要面向应用语言学的研究生和语言测试研究者，后者以实践和应用为重点，主要读者是广大外语教师和教育行政管理人员。本丛书的第一系列围绕语言测试的热点和经典问题展开，主要涉及以下话题：语言测试的效度理论与实证研究、基于任务的语言测试研究、形成性评估研究、ESP 测试研究、动态评估研究、语言测试的评分研究、语言测试计算机自动评分研究、基于语料库的语言测试研究、语言测试的后效研究等。每一本书在保持各自特色的前提下，主要内容包括：①系统介绍相关理论；②本领域的主要研究方法；③实证研究成果分析；④附录：本领域近十年的研究成果。

本丛书的编著者均为语言测试方向的博士和知名学者，相信本丛书的出版将进一步促进国内语言测试研究的发展。本丛书是一个开放的平台，欢迎广大同仁提供自己的新作，欢迎广大读者提出批评与建议。本丛书编著者愿与国内同行一起，为使我国从考试人数大国早日发展成为考试研究大国而努力。

曾用强

2013 年于广州

前 言

计算机技术的发展对语料库与语言测试的发展产生了巨大的影响。得益于计算机技术的普及和进步,机读语料库(machine readable corpus)及其研究和应用开始日益兴盛;对于语言测试,计算机平台带来了巨大的考试变革,如全国首先施行的广东高考中的计算机化英语听说考试。与此同时,语料库对语言测试的发展也有着巨大的影响。而语言测试的发展对计算机技术和语料库的发展也提出了新的要求和预期。因而,三者之间的关系变得错综复杂。局限于篇幅,本书很难把三者都逐一讲解和介绍,因此,本书着眼于这三者的结合点——计算机技术、语料库和语言测试的结合,讨论范围仅限于计算机和语言测试的交叉点、语料库本身和技术方法在语言测试中的应用接口,以及这三者的综合应用。

回顾计算机技术发展的过程,我们注意到,计算机技术在语言测试中的影响力和作用也越来越大(李清华和孔文,2009;Dooey,2008;Chalhoub-Deville,2001;Noijons,1994)。计算机不仅仅用做测试的平台,计算机化语言测试也用于分析传统测试的数据,如信度和效度的研究。在某种意义上,现代的语言测试已经无法离开计算机了。就其根本原因而言,计算机化语言测试日益受到重视与其诸多优点是分不开的。具体说来,这些优点包括了时间节省、信息收集、数据存储、心理测量、多媒体化和标准化(Noijons,1994)。这些优点能够帮助我们更好地实施

语言测试和提高测试的质量。与此同时,大量语料库开始发展后,对语言测试的帮助也是日益增加。就具体应用的途径而言,语料库的应用遍及语言测试的各个环节,主要用途有试题开发、试题优化和试题评估三个方面。对于计算机技术和语料库在语言测试中的应用,本书首先分别讨论其各自对语言测试的影响和作用,然后再探讨这两者的综合应用,最后介绍计算机技术和语料库在语言测试中的具体应用实例:广东高考计算机化英语听说考试的开发和应用,以及基于语义连贯性的计算机自动作文评分系统开发和应用。

本书一共五章和一个附录。第1章给出本书的目的和要点。第2章具体阐述不同形式的计算机化语言测试,其主要内容包括听说读写四种主要考试形式在计算机平台上的实现。第3章则从母语者语料库和学习者语料库两个方面来介绍语料库在计算机化语言测试中的具体应用。第4章讨论计算机技术和语料库在语言测试中的综合应用,提供计算机技术和语料库在语言测试中的综合应用的实例,介绍广东高考计算机化英语听说考试和计算机自动作文评分系统开发和应用。第5章扼要地总结本书的要点和观点。在附录中,提供计算机自动作文评分系统开发和应用中各种统计模型的计算结果。

本书荣幸添列于“现代语言测试与评估丛书”并获得出版资助,本人由衷感谢我的导师、本丛书总编曾用强教授的指导和帮助!此外,本人的师兄李清华教授对本书的写作也给予了大量的帮助和指导,本人在此表示万分谢意!本书的部分内容是基于本人在广东外语外贸大学博士后研究的成果,因此,本人对合作导师刘建达教授和博士后工作站的帮助也深表谢意!

科学出版社致力于扶持学术研究，本人同时对刘彦慧女士为本书编辑所付诸的精力和时间表示感谢！

本书面向在读硕士、博士，语言测试研究者和广大外语教师，既重视理论探讨，也关注实际应用的方法和研究个案；此外，本书还提供本人实践的具体案例，为读者提供将计算机、语料库和语言测试相互结合的实用经验和教训。由于作者水平所限，书中疏漏与讹误之处在所难免，恳请广大同仁不吝批评指正。

作 者

2013 年于广州

目 录

总序

前言

第1章 概述	1
1.1 导言	1
1.2 语言测试和计算机技术	3
1.3 语言测试和语料库发展	9
1.4 本书结构	11
第2章 基于计算机语言测试的主要形式	13
2.1 导言	13
2.2 听力测试	15
2.3 口语测试	24
2.4 阅读测试	39
2.5 写作测试	46
2.5.1 写作工具	47
2.5.2 作文评分员(辅导员)	52
2.6 小结	56
第3章 语料库在语言测试发展中的作用	57
3.1 导言	57

3.2 学习者语料库的作用	59
3.3 母语语料库的应用	69
3.4 小结	83
第4章 计算机技术和语料库的综合应用实例	84
4.1 导言	84
4.2 广东省高考计算机化英语听说考试	84
4.2.1 考试的发展过程和形式	84
4.2.2 试题开发和效度验证	90
4.2.3 小结	100
4.3 基于连贯性分析的计算机自动作文评测系统开发和应用	102
4.3.1 引言	102
4.3.2 研究背景和问题	103
4.3.3 研究问题	109
4.3.4 CTutor 的开发	109
4.3.5 实验设计和数据	116
4.3.6 结果	118
4.3.7 讨论和结语	120
4.4 小结	121
第5章 计算机和语料库的应用前景	123
5.1 语言教学和测试的新概念和新应用	123
5.2 测试发展对计算机和语料库的新需求	129
5.3 结语	132
参考文献	133

附录 CTutor 各个统计模型计算结果	152
NMET Regression Curve Estimation Results	152
CET-4 Regression Curve Estimation Results	159
CET-6 Regression Curve Estimation	166
TEM 4 Regression Curve Estimation	173
TEM 8 Regression Curve Estimation	180

第 1 章

概 述

1.1 导 言

计算机技术、语料库与语言测试三者之间的关系很难用一两句话概括清楚，因为这三者本身都是一门独立发展的学科或者产业。本书把这三者放到一起讨论是因为前两者都可以在语言测试中结合起来，即计算机技术和语料库都能够促进语言测试的进步和发展。在进入具体的讨论之前，有必要澄清本书中所涉及的主要概念。在本书中，语言测试主要是指对中国外语学习者所学外语能力的测量和评估。计算机技术这个概念比较笼统，包括了软件、硬件和网络。计算机自适应考试和基于网络的语言测试，在本书中也视为计算机化语言测试的实现方式。语料库主要指的是现代机读语料库，也就是说这类语料库能够在计算机平台上存储和查询。本书的讨论范围也限于计算机和语言测试的交叉点，语料库本身和技术方法在语言测试中的应用接口，以及这三者的综合应用。

需要说明的是，本书的内容是部分基于本人的博士后研究工作中的理论探讨和思考。因此，本书的目的不是说服读者在语言测试中应用计

算机技术和语料库方法，也无意声称基于计算机和语料库的语言测试能够优于传统的笔纸测试。本书是为了让读者了解计算机和语料库对语言测试的帮助和作用，也介绍这些新的技术方法可能存在的局限性。因此，本书尽量本着客观的态度来介绍和说明计算机技术和语料库方法在语言测试中的作用和影响。

把计算机技术和语料库技术引入语言测试的目的是为了提高和改进语言测试的信度和效度，即更好地实现语言测试的目的。而语言测试的目的就是为了了解和判断一个人能够如何有效地使用其所学的语言。这种了解和判断在不同的环境中有不同的用途(use of language tests)。比如，在日常教学中，语言测试帮助教师或者学生分析学生的学习进度和存在的问题；在入学考试中，它就成为筛选优秀学生的一个方法和途径；在以移民为目的的语言考试中，语言测试成为预测申请人融入当地社会可能性的一个分析手段。语言测试的用途本身更多地和社会文化因素结合在一起，成为语言测试研究领域的一个重要组成部分。本书不对语言测试的具体用途展开讨论，而是探讨计算机技术和语料库方法如何帮助语言测试研究者和开发者改进测试方法从而提高测试的信度和效度。

语言测试的信度和效度是一个重要而又复杂的话题，但是提高信度和效度的本质是更好地收集关于学习者语言知识和语言使用能力的相关信息(Chapelle, Enright & Jamieson, 2008; Chapelle, 1999)，进而依据不同的考试用途而做出尽可能准确的研判。简单说来，语言测试就是一个信息搜集的过程。关于获取何种信息，现代语言测试则更多地关注学习者语言使用过程中的表现以及涉及相关社会文化知识的交际能力(Myers, 2002)。而且，通过测试所获取的信息是关于受试者会什么而

不是不会什么。由于计算机技术和语料库方法的介入,这一信息的收集过程变得更加科学和准确。相对于传统的语言测试方法,计算机技术和语料库方法可以帮助语言测试的开发和研究人员利用科技手段更好地了解受试者在语言使用过程中的表现。简单地说,现代的计算机技术和语料库方法可以帮助语言测试的开发和研究人员在特定的时间和空间范围内,获取更多关于语言使用的相关信息。而这些信息可以帮助我们更加全面地判断受试者语言使用能力及其所掌握的相关社会文化交际能力。因此,语言测试的开发和研究人员通过计算机技术和语料库的方法来更加全面和详尽地获取受试者语言使用信息,从而对受试者有更为准确的判断和了解。同时,把正确的判断结论和相应的语言测试用途结合起来,提高测试的效度和信度。

1.2 语言测试和计算机技术

计算机技术从很早就开始在语言测试中应用,伴随着技术的进步,计算机技术在语言测试中的影响力和作用也越来越大(李清华和孔文, 2009; Dooley, 2008; Chalhoub-Deville, 2001; Noijons, 1994;)。计算机不仅仅用做测试的平台,计算机化语言测试也用于分析传统测试的数据,如信度和效度的研究。在某种意义上,现代的语言测试已经无法离开计算机了。

虽然计算机在语言测试中的地位越来越重要,但是我们也应当注意到其局限性。例如,尽管计算机化语言测试有诸多方面的优势,但是同时熟悉计算机技术和语言测试的专家、考试开发人员比较少。甚至还有

人误以为计算机化测试就是把传统的笔纸测试电子化即可，却不了解以计算机为工作平台的语言测试有其自身的特点和要求。这些特点体现在一个考试制作和使用的各个阶段(Noijons, 1994)，主要是在考试制作、实施和评分三个环节。

在考试制作的环节中，可以是计算机自动产生试题或者利用计算机来辅助制作考试试题。前者既可以是完形填空题的自动生成，也可以是利用人工筛选的试题形成试题库，然后依据考试需要，计算机根据难度系数和区分度等数字指标自动生成考试试题。在计算机辅助试题生成中，通过机器来分析试题素材的一些语言特征，如可读性、词汇难度和语句复杂度等，从而帮助考试制作人员来选择试题原材料完成试题制作的过程。

考试实施环节中，计算机辅助语言测试多指利用计算机或者以计算机为载体的网络作为考试实施的平台。换言之，计算机和受试者实时交互来完成考试。平时我们所说的计算机化语言测试，就是指在考试实施环节实现计算机和受试者的实时互动。很多涉及计算机化语言测试的问题都和这个环节有关系，如由于计算机水平不一致而产生的考试公平性问题、电脑程序设计中采用的布局和考试流程安排等。其实这些问题不难理解，在计算机比传统笔纸考试为语言测试提供更多灵活性的同时，也引入了更多影响语言测试信度和效度的变量，也就意味着考试本身出现问题的概率也会相应增加，与此同时，我们所关心或所需要考虑的因素也增多。因此，很多关于计算机化语言测试的研究都是围绕这个环节来展开的。

在第三个环节中, 考试评分可以在考试实施过程中自动完成, 如选择题、完形填空题等题型, 但是有些试题类型由于技术难度等原因不得不在考试实施完成之后才进行, 如写作。把各个部分的得分汇总形成一个综合性的总分, 也还是在考试实施的环节之后完成的。除了分数的生成和统计之外, 把考试结果存储到计算机中也是一个重要的步骤。结合受试者的年龄、性别、社会文化背景和以往学习记录等其他相关信息, 考试的开发或研究人员可以进一步研判受试者在考试中的表现, 也可以利用这些数据来评估考试本身。

就总体而言, 计算机化语言测试日益受到重视与其诸多优点是分不开的。这些优点包括了时间节省、信息收集、数据存储、心理测量、多媒体化和标准化(Noijons, 1994)。

从理论角度而言, 计算机化语言测试在节约考试时间方面是有优势的。计算机平台比传统的笔纸考试平台能够在同等时间内获取更多的信息量, 从而能够更加迅速地对受试者的语言使用水平做出正确的判断。并且, 计算机化语言测试能够实现自适应考试(adaptive testing/computerized adaptive testing), 即依据受试者不同的能力给出不同的试题。也就是说, 不同能力水平的受试者在同一考试中会接触到不同的试题。这种自适应考试能够大大缩短考试所需的时间。在考试过程本身之外, 试题的制作、存储、传递和评分实现计算机化后, 也能够相对于传统考试节约大量时间。

计算机在信息收集方面有巨大的优势。从考试开始, 受试者所有的行为都可以记录下来, 如受试者在每个试题部分使用的时间、受试者对

答案的更改和反复等。随着技术的进步,结合带有双摄像头的计算机和相关的软件,可以把受试者的眼球移动轨迹记录下来。这些信息虽然没有和语言使用水平直接联系,但是也可以帮助我们对受试者的语言使用能力做出更加精准的判断。

高速和大容量的数据存储是计算机化语言测试另一个重要的优势。借助于计算机本身的硬盘和外设的光盘、外置硬盘等存储介质,计算机可以实现理论上的无穷尽海量数据存储。不同于传统的语言测试,其数据大部分依赖于受试者作答的试卷为载体。在计算机化语言测试中,特别是在近些年实现的云计算(cloud computation)中(Piovesan, Hoff do Amaral, Arenhardt & Medina, 2012),数据存储不仅速度高、容量大而且网络化、与平台无关,即在一种操作系统中生成的数据通过云计算平台可以为其他操作系统所共享(Cubillo, Martín & Castro, 2011)。比如,在 Windows 计算机上完成的考试,通过云计算平台可以在 iOS 系统、UNIX 系统和 Linux 系统上处理和分析。

借助计算机平台数据处理的便利性,心理测量在计算机化语言测试中较为容易实现。例如,得益于大型试题库和计算机的实时处理能力,我们可以方便而又及时地计算出受试者所适合的试题难度值。计算机化语言测试可以通过相对少量的试题对受试者语言使用能力有一个准确的评估,而无须通过大量的试题。在计算机平台上的自适应考试就是对心理测量优势的充分利用。

有了计算机的多媒体功能,从平面文字到声音图像,计算机化语言测试能够为受试者创造一个更加逼真的语言使用环境,这是传统型的笔